

Time domain methods for speech processing

- **Positive definite matrix**
- *Definition of a positive definite matrix*
- A symmetric matrix A is said to be *positive semi-definite* if, for any non 0 vector x :
- $x'Ax \geq 0$
-

- **Properties of positive definite matrices**
- A positive (semi-)definite matrix enjoys all the good properties of symmetric matrices. In addition, its positive ness brings about new important properties, of which we list a few with no attempt at being systematic :
- * A positive definite matrix is nonsingular, and therefore has an inverse, which turns out to be also positive definite.
- * Any principal square submatrix of a positive definite matrix is positive definite.

- A psd matrix has many non symmetric square roots, among which the one obtained by the Cholesky factorization is of particular interest

Time Domain Methods in Speech

- **Short-Time Energy**

$$E = \sum_{m=-\infty}^{\infty} x^2(m)$$

this is the long term definition of signal energy

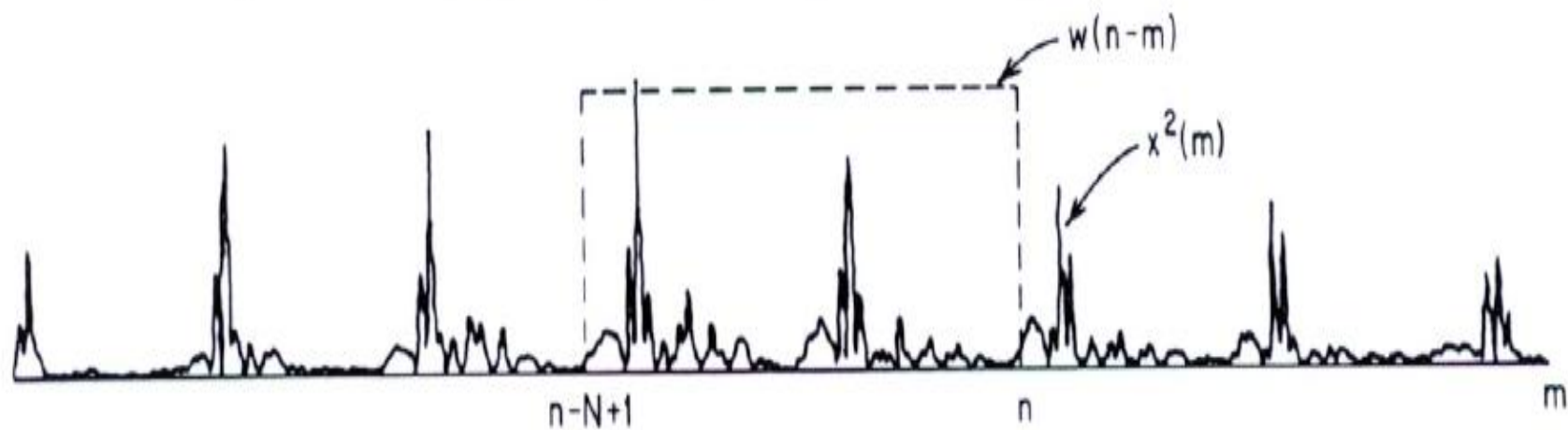
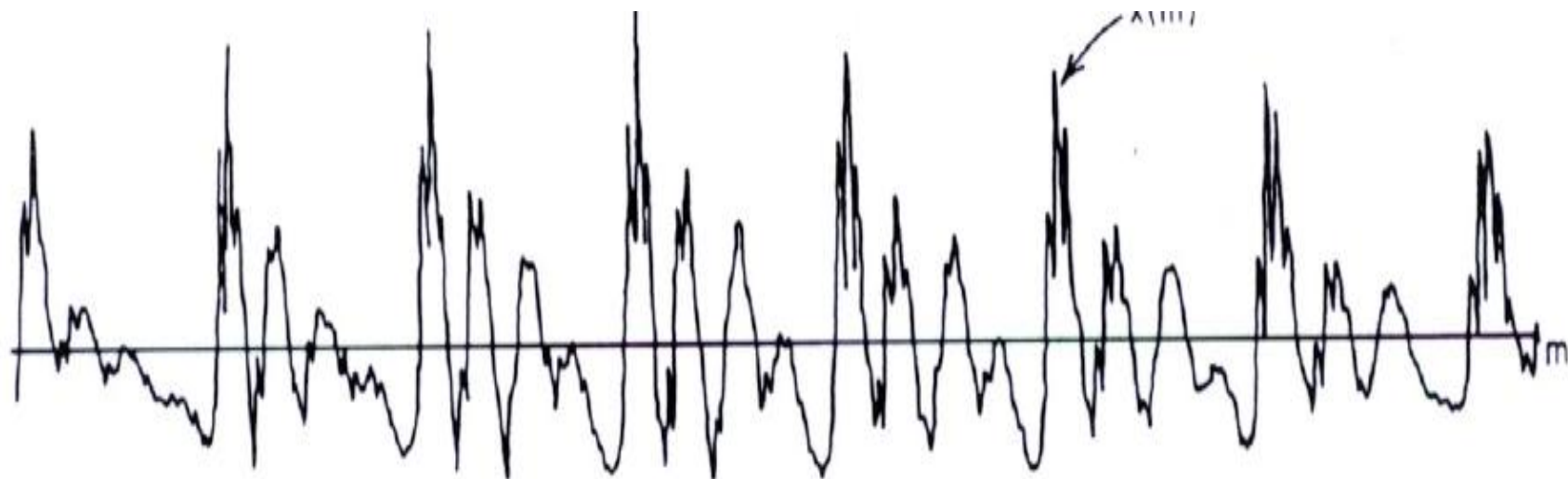
there is little or no utility of this definition for time-varying signals

$$E_n = \sum_{m=n-N+1}^n x^2(m) = x^2(n-N+1) + \dots + x^2(n)$$

short-time energy in vicinity of time n

$$T[x] = x^2$$

$$w(n) = 1 \quad 0 \leq n \leq N-1$$
$$= 0 \quad \text{otherwise}$$



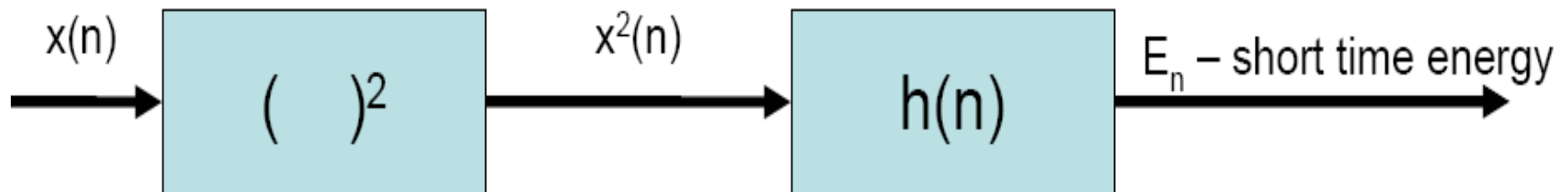
Short-Time Energy

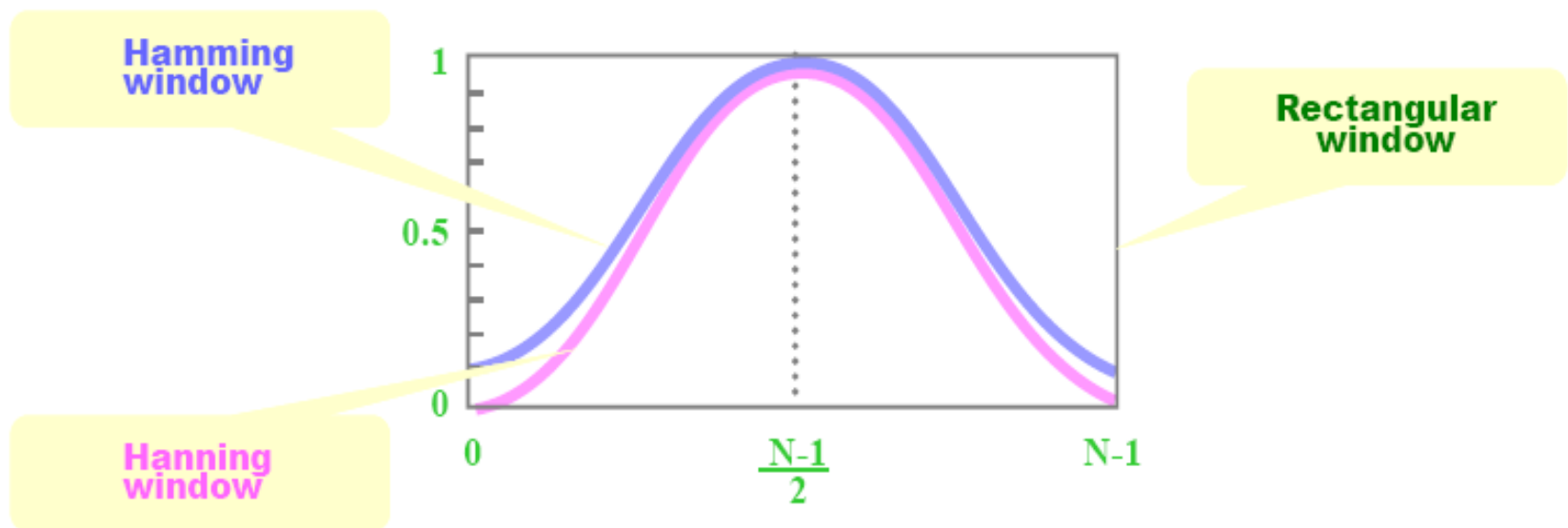
$$E_n = \sum_{m=-\infty}^{\infty} [x(m) w(n-m)]^2 \text{ (sum of squares of portion of signal)}$$

-- concentrates measurement at sample n , using weighting $w(n-m)$

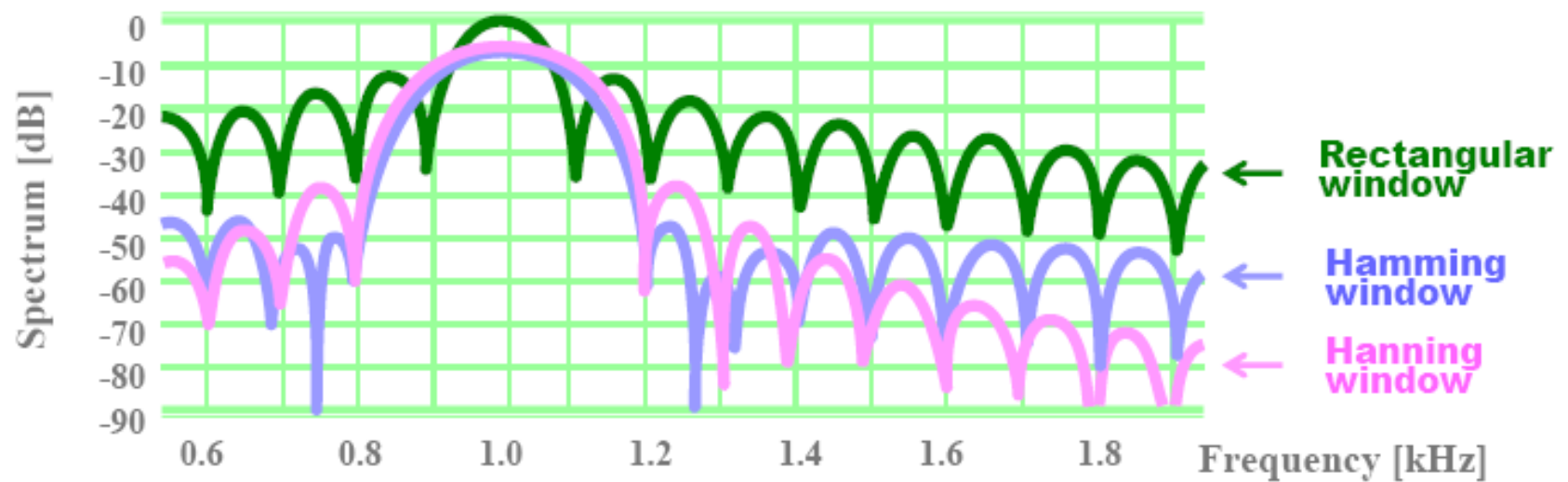
$$E_n = \sum_{m=-\infty}^{\infty} x^2(m) w^2(n-m) = \sum_{m=-\infty}^{\infty} x^2(m) h(n-m)$$

$$h(n) = w^2(n)$$





(a) Major window functions



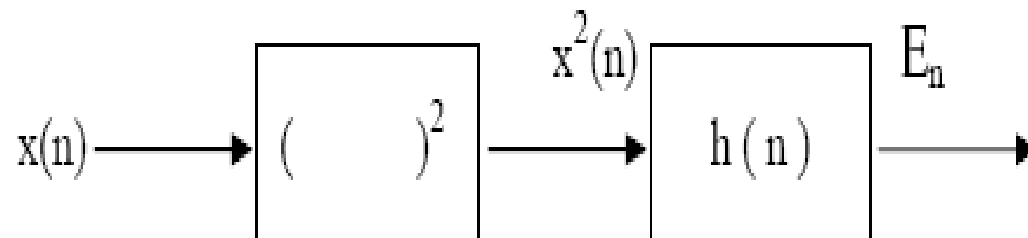
b) The spectrum for the 10 periods of a 1-kHz sinusoidal wave extracted using each of the windows

$$Q_n = \sum_{m=-\alpha}^{\alpha} T[x(m)] w(n-m)$$

$$E_n = \sum_{m=-\alpha}^{\alpha} [x(m) w(n-m)]^2$$

$$E_n = \sum_{m=-\alpha}^{\alpha} x^2(m) h(n-m)$$

$$h(n) = w^2(n)$$



- Increasing the Length N , simply decrease the band width.
- If N is too small, i.e. on the order of a pitch period or less, E_n will fluctuate very rapidly depending on exact details of the waveform.
- If N is too large, i.e. on the order of several pitch periods, E_n will change very slowly and this will not adequately reflect the changing properties of the speech signal

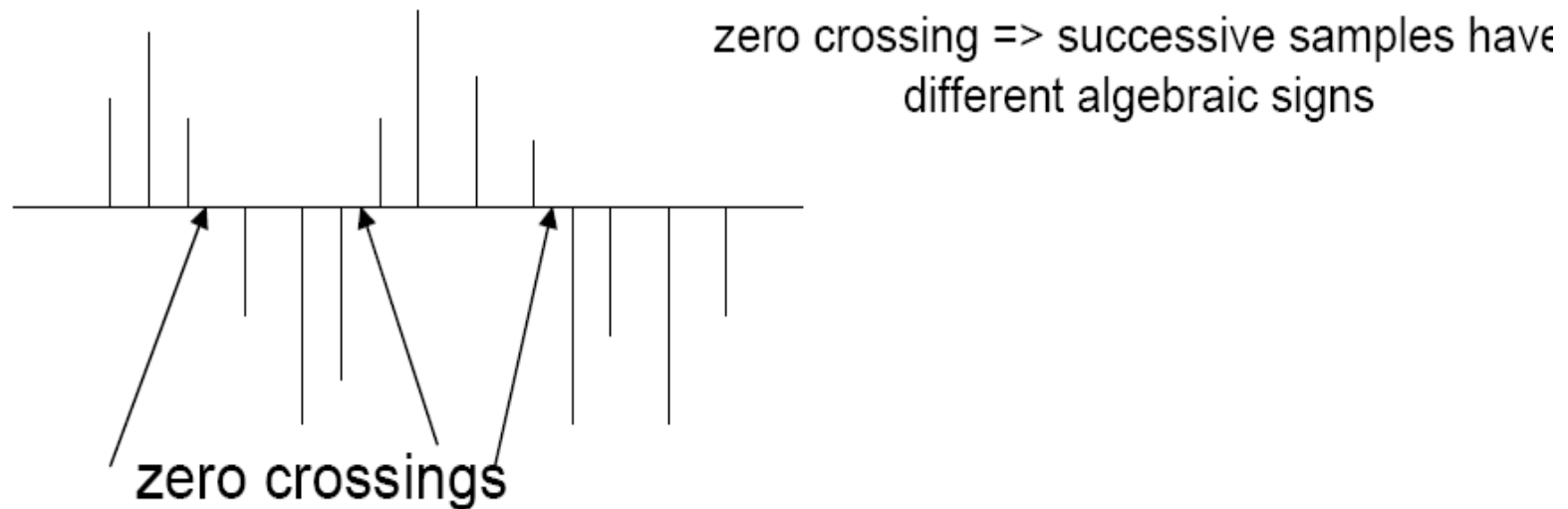
Average Magnitude Difference Function

Difficulty with short-time energy function is that, it is very sensitive to large signal levels, emphasizing large sample to sample variations of $x(n)$.

sum of absolute values of the signal is computed instead of the sum of squares

$$M_n = \sum_{m=-\alpha}^{\alpha} |x(m)| w(n-m)$$

Short-Time Average ZC Rate



- zero crossing rate is a simple measure of the 'frequency content' of a signal—especially true for narrowband signals (e.g., sinusoids)
- sinusoid at frequency F_0 with sampling rate F_S has F_S/F_0 samples per cycle with two zero crossings per cycle, giving an average zero crossing rate of

$$Z=2F_0/F_S \text{ crossings/sample}$$

Zero Crossing Rate

Number of zero-crossings is related to a sinusoid's frequency. A sinusoid of frequency n Hz has $2n$ zero-crossings per second

Consider a window of size N samples with Z zero-crossings and a sampling frequency f_s .

Then number of zero-crossings per second = $Z / (N / f_s)$

and we define the zero-crossing rate as

$$\text{ZCR} = Z f_s / (2N) \text{ Hz}$$

e.g. $Z=10$, $N=200$, $f_s=16000$ gives

$\text{ZCR} = 400$ Hz which is close to first formant frequency.

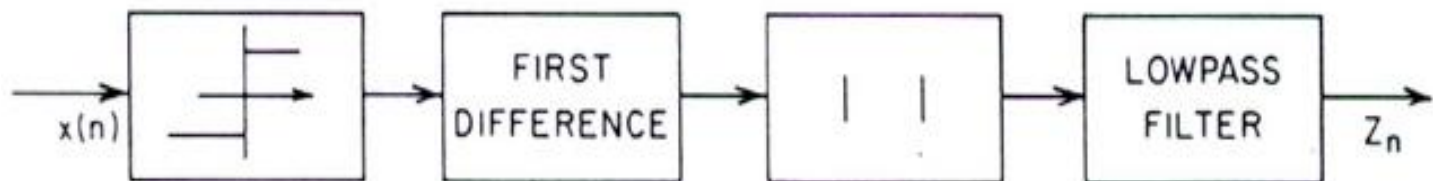
ZC Rate Definitions

$$Z_n = \sum_{m=-\infty}^{\infty} |\operatorname{sgn}[x(m)] - \operatorname{sgn}[x(m-1)]| w(n-m)$$

$$\begin{aligned} \operatorname{sgn}[x(n)] &= 1 & x(n) &\geq 0 \\ &= -1 & x(n) &< 0 \end{aligned}$$

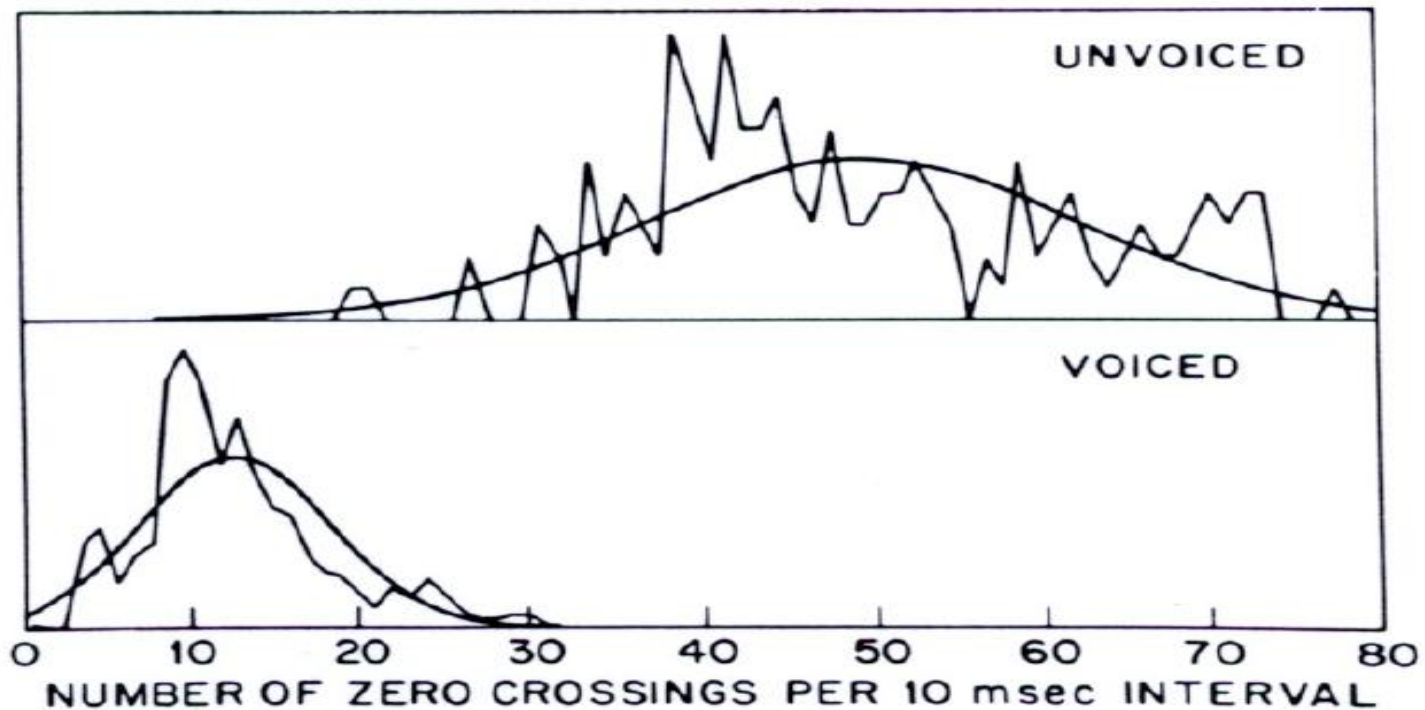
- simple window, rectangular with

$$\begin{aligned} w(n) &= 1/(2N) & 0 \leq n \leq N-1 \\ &= 0 & \text{otherwise} \end{aligned}$$



- $\Rightarrow$ Zero crossings is high $\Rightarrow$ Sp
signal is unvoiced
- $\Rightarrow$ Zero crossings is low $\Rightarrow$ SP
signal is voiced

histogram of average zero crossing rates (averages
over 10 m sec.)
for both voiced and unvoiced sp.

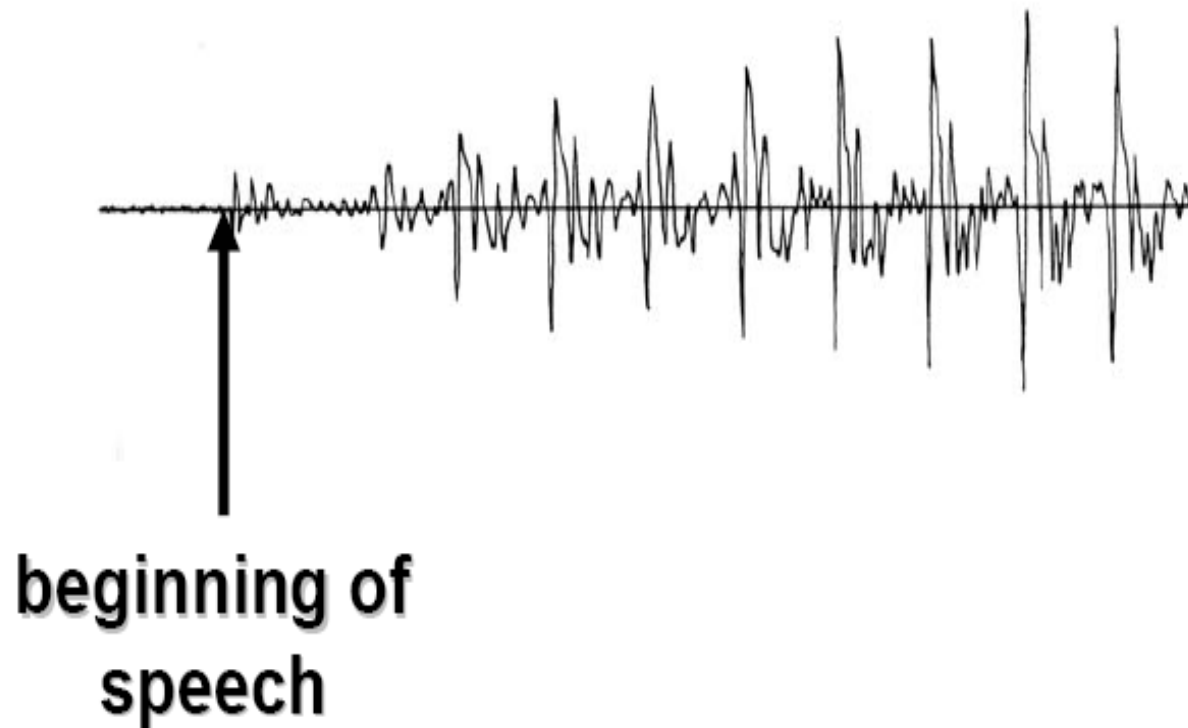


End Point Detection

The problem of locating the beginning and end of a speech utterance in a background noise.

automatic recognition of isolated words, it is essential to locate the regions of a speech signal that corresponds to each word.

key problem in speech processing is locating accurately the beginning and end of a speech utterance in noise/background signal



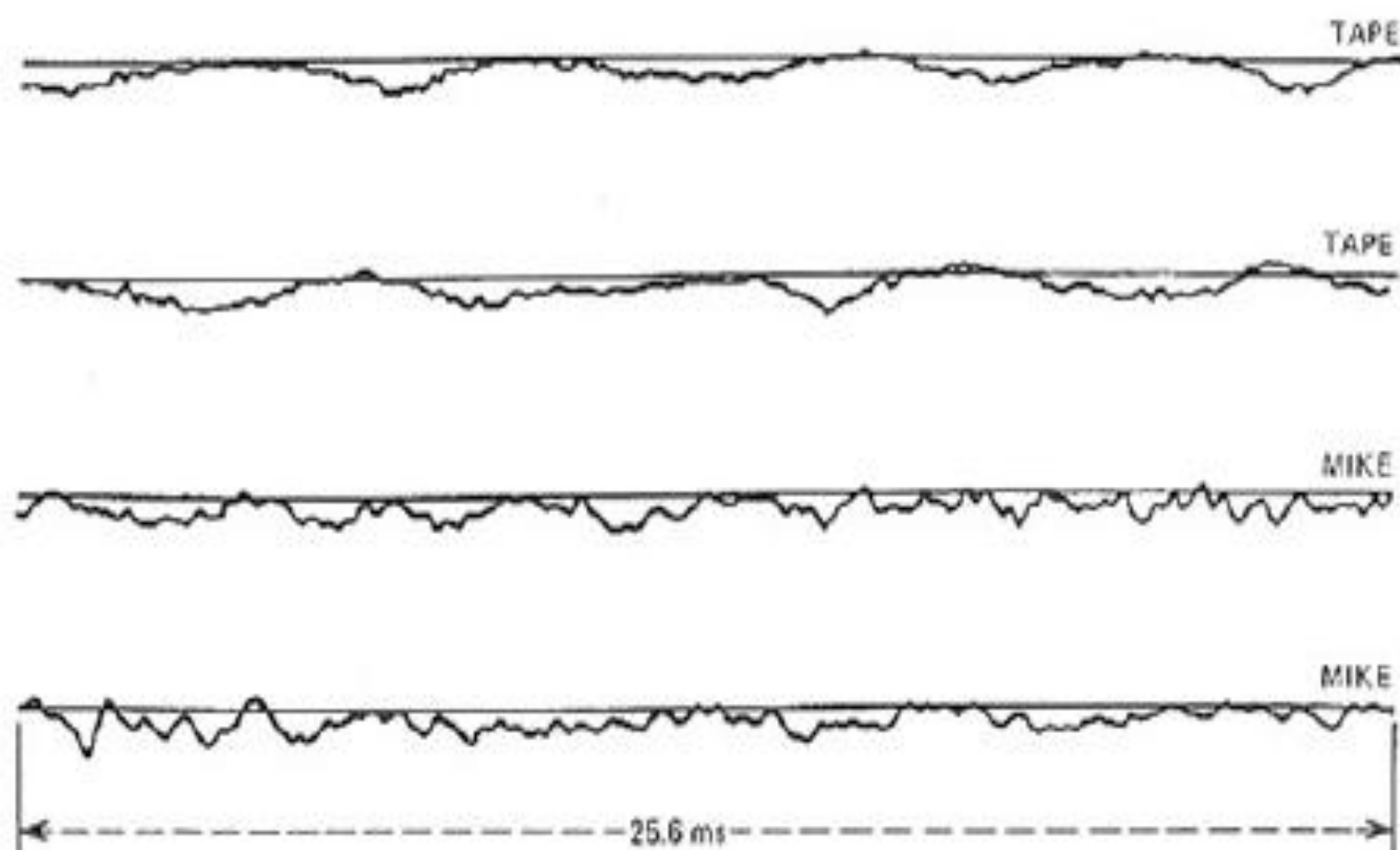


Fig. 1—Acoustic waveforms for the silences from tape and microphone.

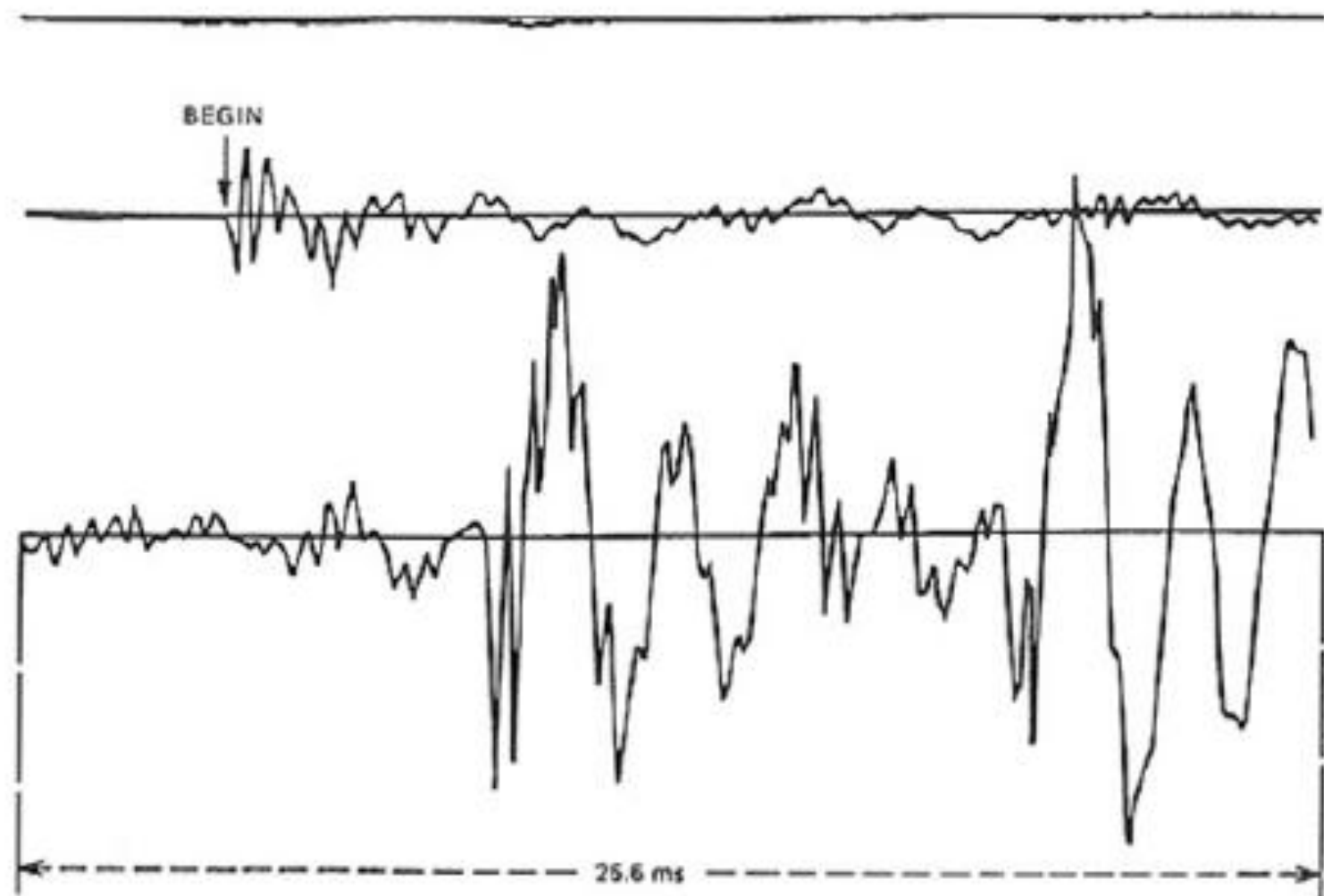


Fig. 3—Waveform for the beginning of the word "eight."

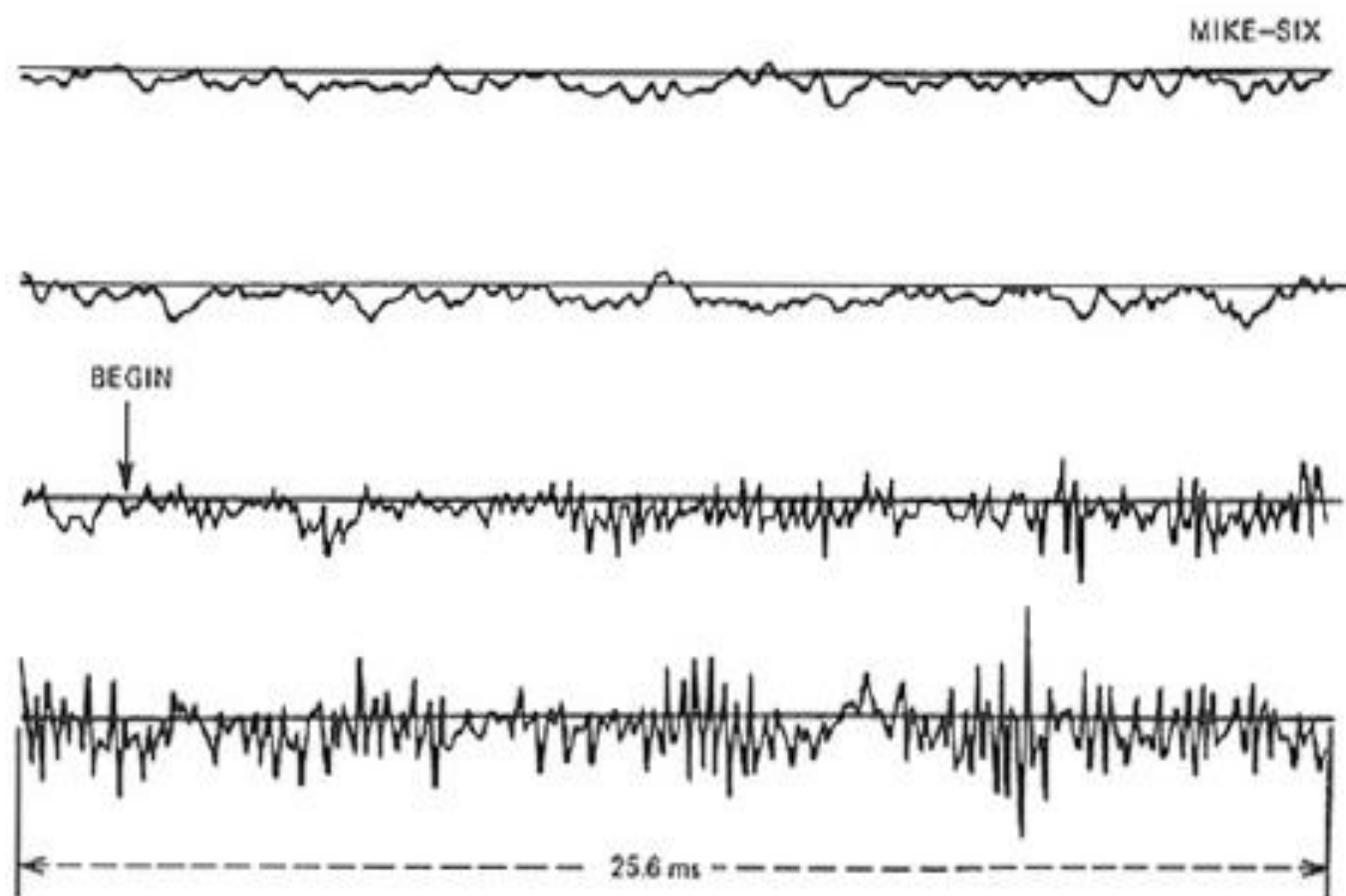


Fig. 4—Waveform for the beginning of the word "six."

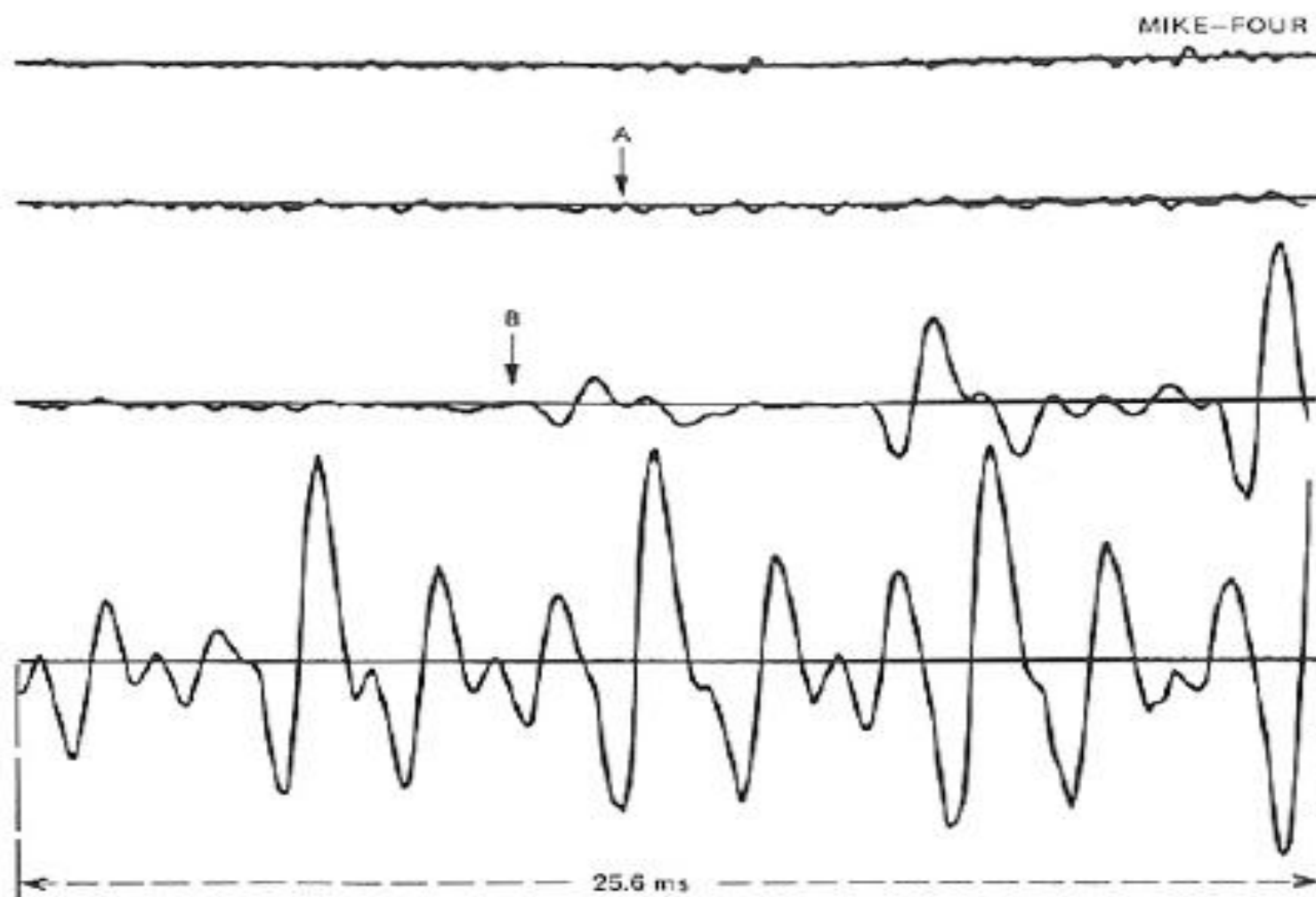


Fig. 5—Waveform for the beginning of the word "four."

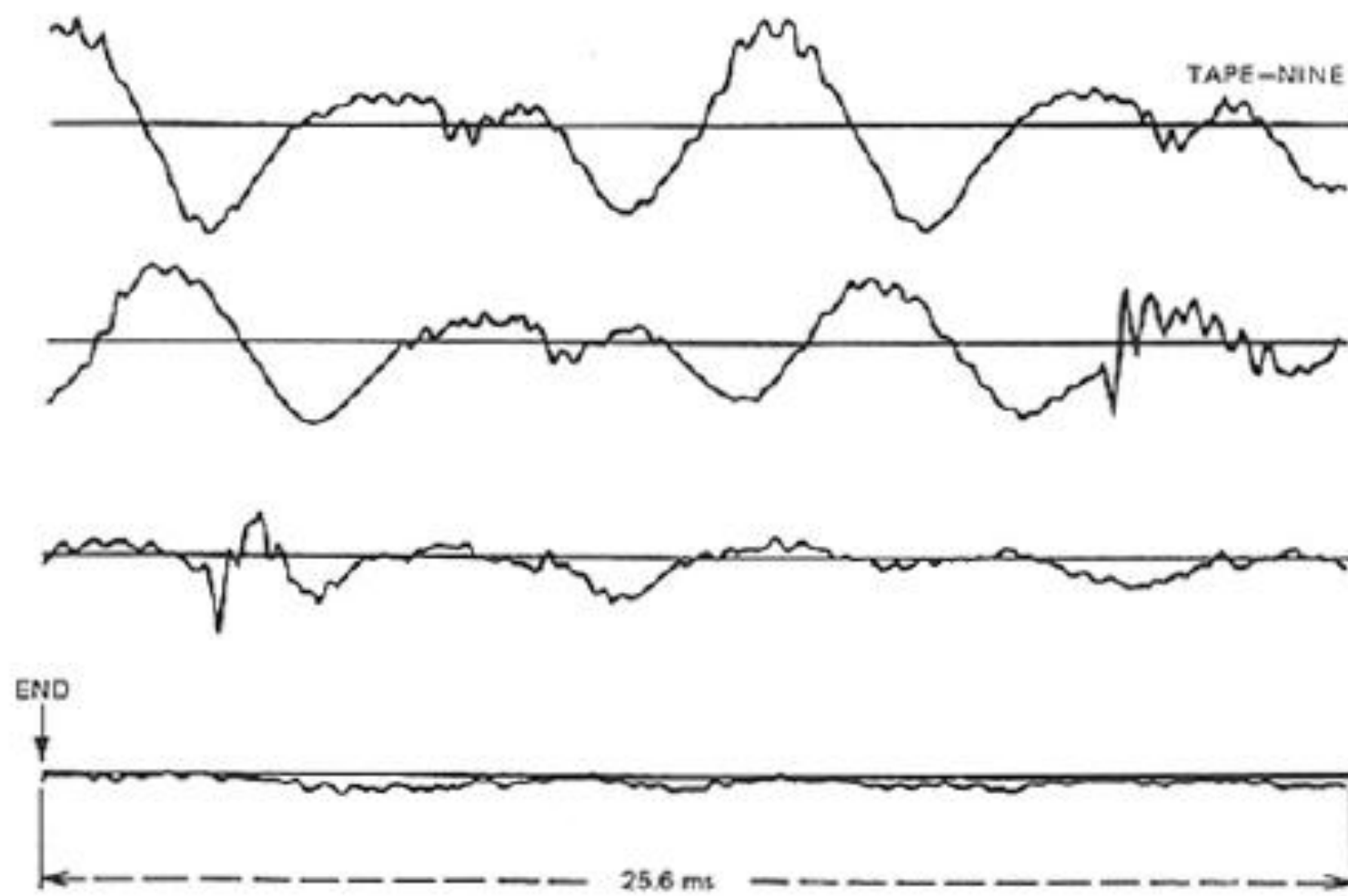


Fig. 7—Waveform for the end of the word "nine."

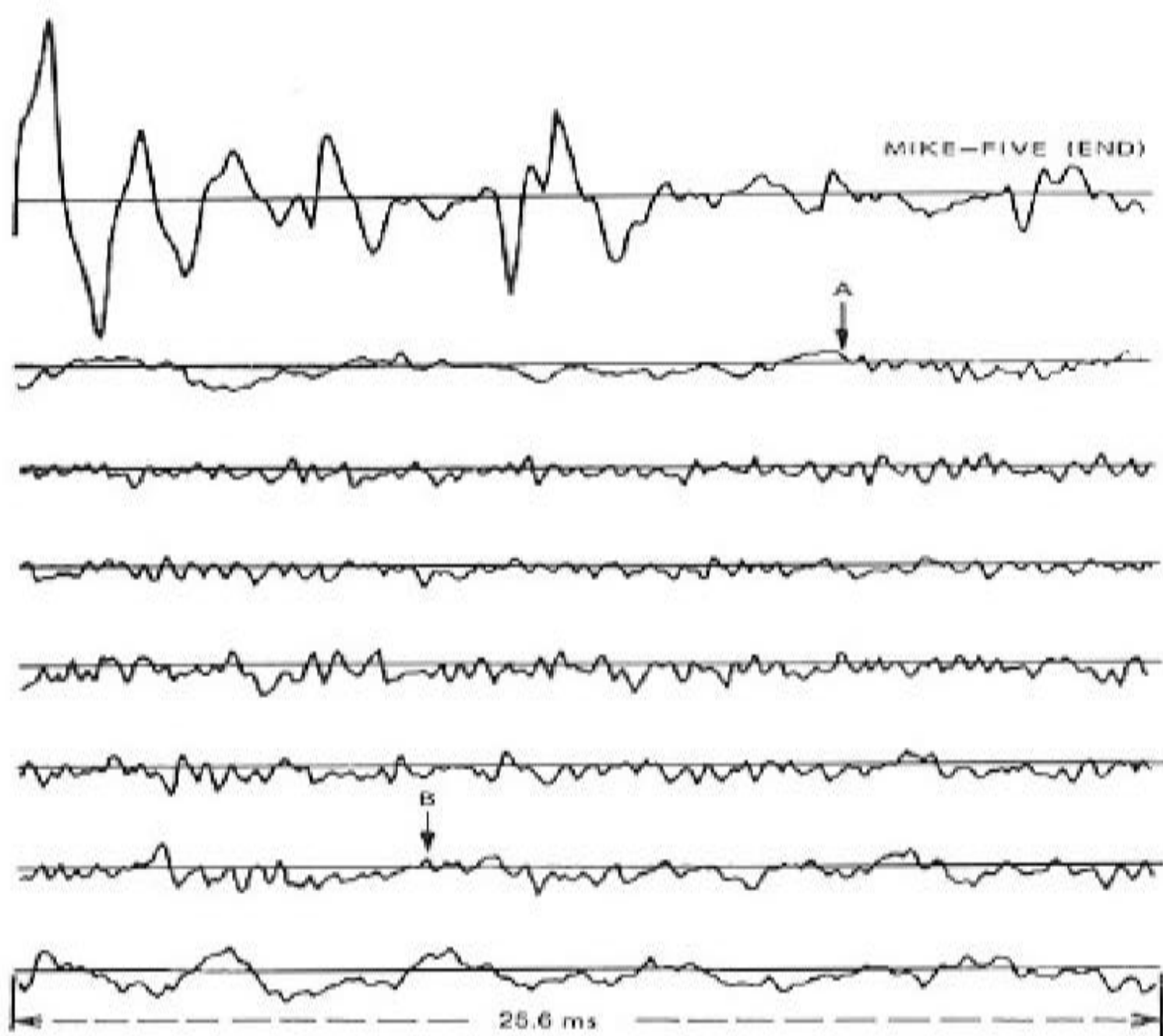


Fig. 6—Waveform for the end of the word "five."

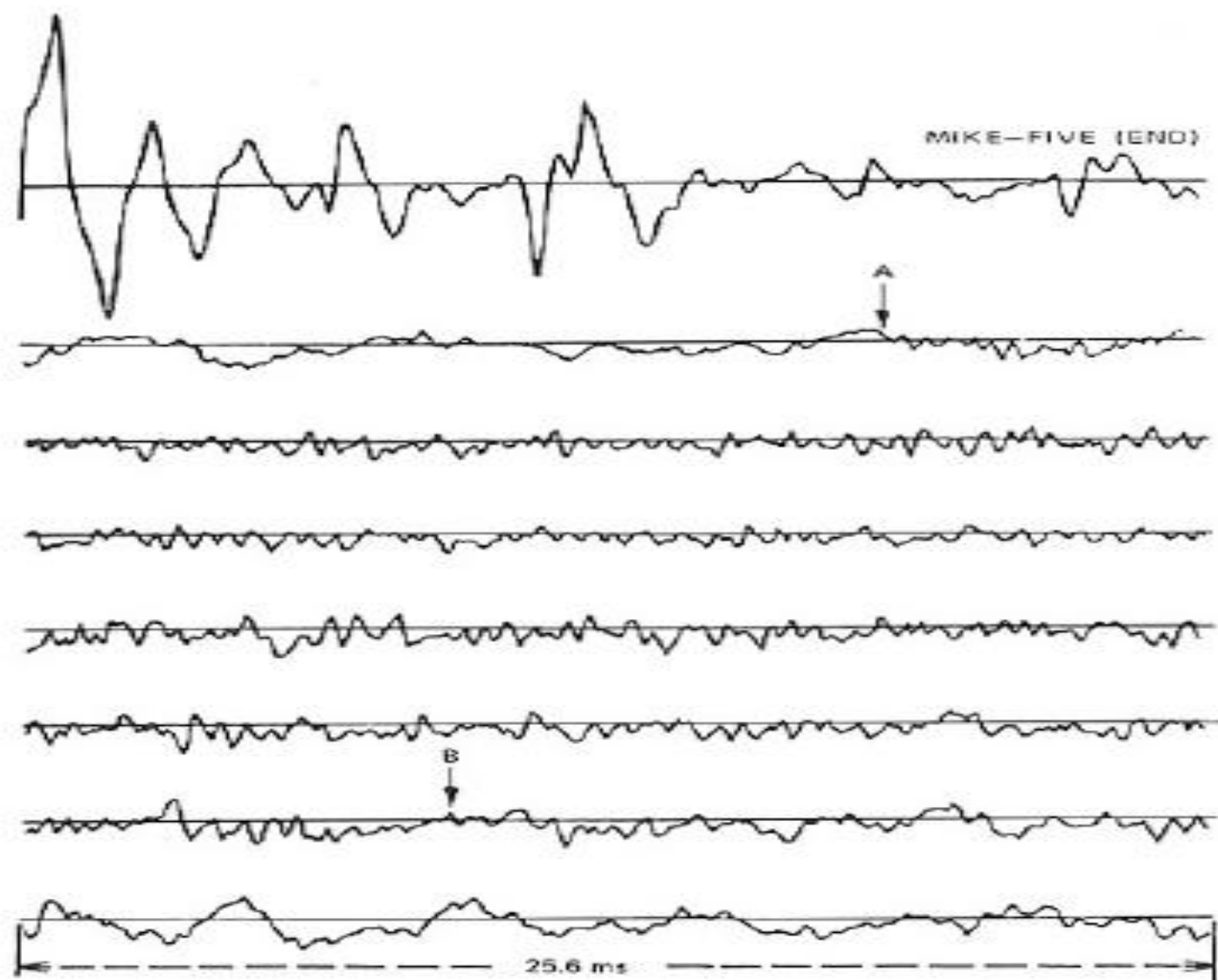


Fig. 6—Waveform for the end of the word "five."

- It is difficult to locate the beginning and end of an utterance if there are :
- Weak fricatives (/ f /, / th / and / h /) at the beginning or end.
- Weak plosive bursts (/ p /, / t / and / k / at the beginning or end.
- Nasals at the end.
- Voiced fricatives which become devoiced at the end of words.
- Trailing off of vowel sounds at the end of utterance.

- **End Point Detection** Algorithm
- 1. Detect beginning and ending of speech intervals using short time energy and Short-time zero crossings.
- 2. Find major concentration of signal (guaranteed to be speech) using region of Signal energy around maximum value of short time energy.
- 3. Refine region of concentration of speech using reasonably tight short-time energy thresholds that separate speech from backgrounds but may fail to find weak fricatives, low level Nasals etc.
- 4. Refine end point estimates using zero crossing information outside intervals identified from energy concentrations-based on zero crossing rates commensurate with voiced speech.

$$IZCT = \text{MIN}(IF, \overline{IZC} + 2\sigma_{IZC}).$$

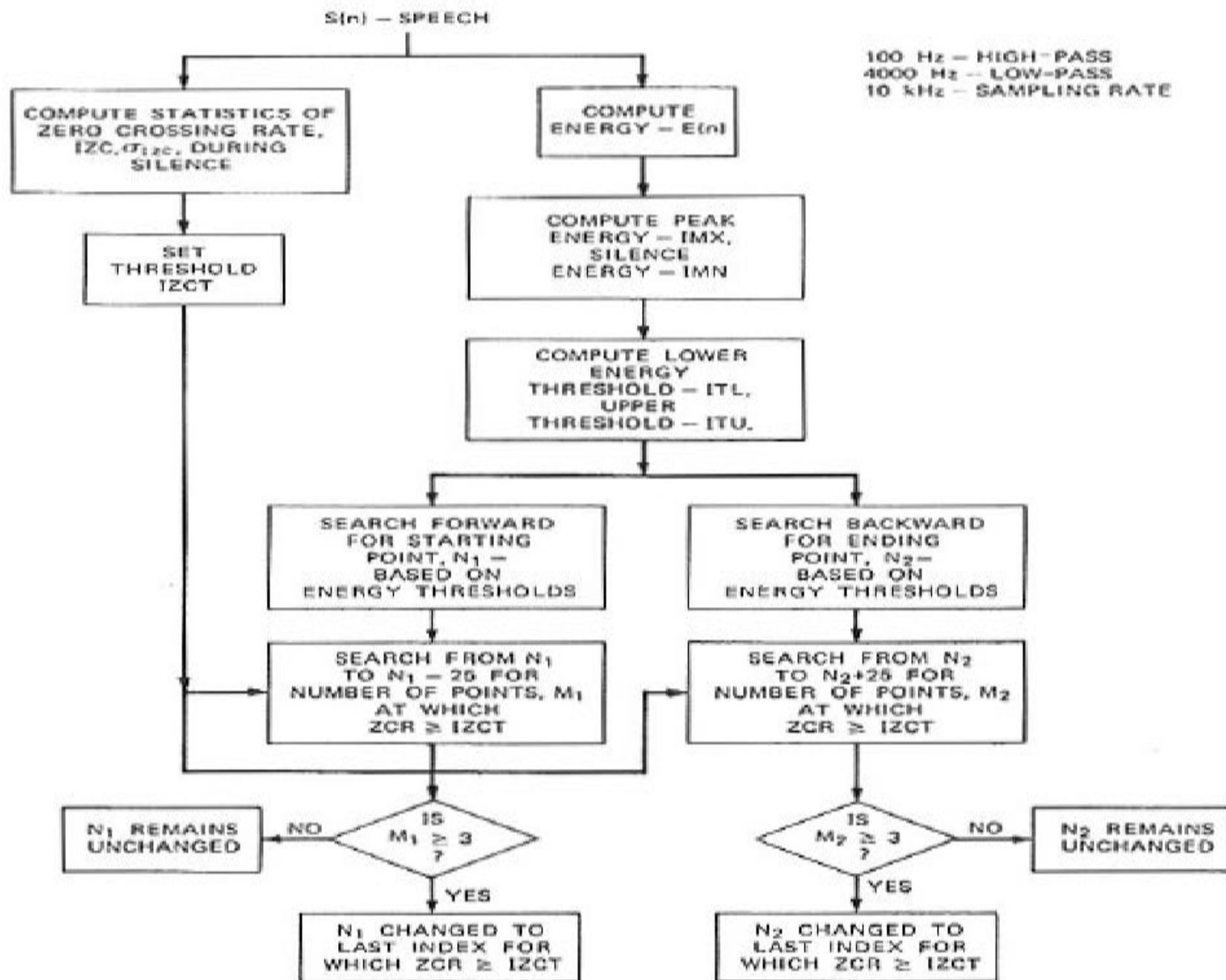
The energy function for the entire interval, $E(n)$, is then computed. The peak energy, IMX , and the silence energy, IMN , are used to set two thresholds, ITL and ITU , according to the rule

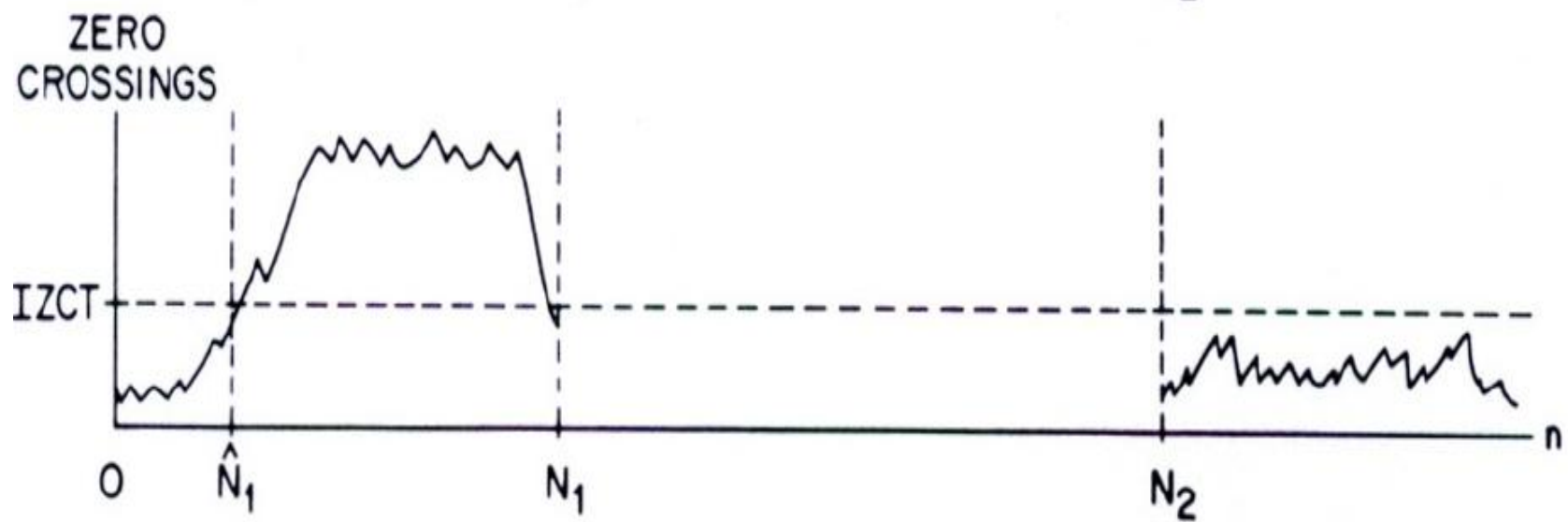
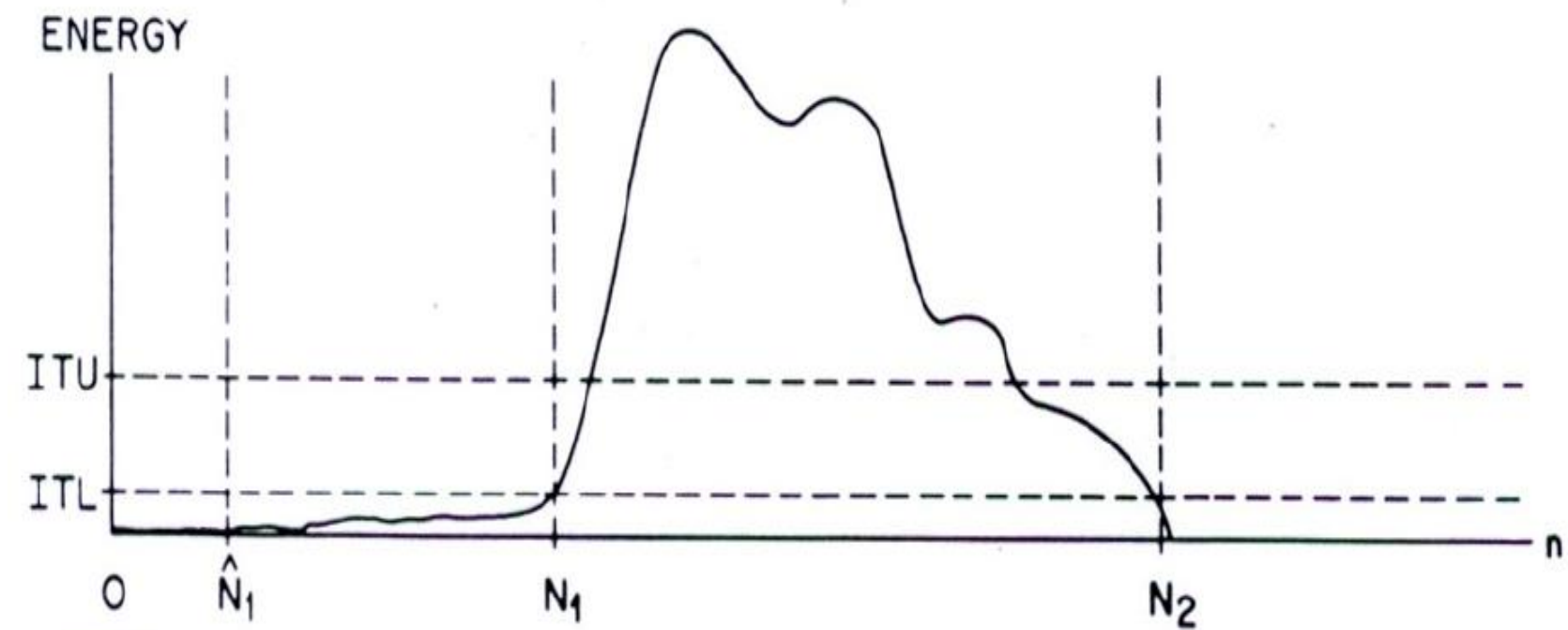
$$I1 = 0.03*(IMX - IMN) + IMN \quad (3)$$

$$I2 = 4*IMN \quad (4)$$

$$ITL = \text{MIN}(I1, I2) \quad (5)$$

$$ITU = 5*ITL. \quad (6)$$





End point Detection using short-time Measures

Algorithm for end point Detection

Complete mean and variance of E_n and Z_n for first 100 m sec. of signal (assuming no speech in this interval).

Determine maximum value of E_n for entire recording.

Compute E_n thresholds based on steps (1) and (2) .

eg: take some percentage of the parts over the entire interval. Use threshold for zero crossings based on Z_c distribution for unvoiced speech.

- Find an interval of E_n that exceeds a high threshold ITU.
- Find a putative (supposed) starting point ($N1$) where E_n crosses ITL from below.
- Find a putative ending point ($N2$) where E_n crosses ITL from above.
- Move backwards from $N1$ by comparing Z_n to IZCT, and find the first point where Z_n exceeds IZCT;
- similarly move forward from $N2$ by comparing Z_n to IZCT and finding last point where Z_n exceeds IZCT.